General Utterance-Level Feature Extraction for Classifying Crying Sounds, Atypical & Self-Assessed Affect and Heart Beats

Gábor Gosztolya[1], *Tamás Grósz*[1,2], *László Tóth*[2]

[1] MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[2] Institute of Informatics, University of Szeged, Hungary
{ ggabor, groszt, tothl } @ inf.u-szeged.hu

Abstract

In the area of computational paralinguistics, there is a growing need for general techniques that can be applied in a variety of tasks, and which can be easily realized using standard and publicly available tools. In our contribution to the 2018 Interspeech Computational Paralinguistic Challenge (ComParE), we test four general ways of extracting features. Besides the standard ComParE feature set consisting of 6373 diverse attributes, we experiment with two variations of Bag-of-Audio-Words representations, and define a simple feature set inspired by Gaussian Mixture Models. Our results indicate that the UAR scores obtained via the different approaches vary among the tasks. In our view, this is mainly because most feature sets tested were local by nature, and they could not properly represent the utterances of the Atypical Affect and Self-Assessed Affect Sub-Challenges. On the Crying Sub-Challenge, however, a simple combination of all four feature sets proved to be effective.
Index Terms: computational paralinguistics, ComParE 2018, classification, posterior estimates, posterior calibration

1. Introduction

Computational paralinguistics consists of a great variety of tasks, including emotion detection [1, 2], determining the amount of cognitive load [3], whether the speaker has some kind of physical (e.g. cold [4, 5]) or mental illness (e.g. depression [6], Parkinson's disease [7] or Alzheimer's disease [8]). Availability and standardization of datasets is facilitated by the Interspeech Computational Paralinguistic Challenge (ComParE), held annually at the Interspeech conference.

Although state-of-the-art paralinguistic performance is usually achieved by incorporating task-specific features or techniques (such as determining the duration when several people are speaking at the same time for conflict intensity estimation [9, 10])), there is a growing need for general approaches which perform well for several different computational paralinguistic tasks. This need can be seen even in the baselines of the ComParE Challenges: in 2017 and 2018 (see [11] and [12]), baseline systems did not exclusively rely on the traditional, 6373-sized feature set like the Challenges of the preceding years, but incorporated BoAW extraction [13], end-to-end learning, and from this year, sequence-to-sequence autoencoders as well.

Due to this, in our contribution to the 2018 ComParE Challenge [12], we experiment with different feature sets designed with general paralinguistic applicability in mind. One is the classic, 6373-sized paralinguistic feature set developed by Schuller et al. [14]. The second and third feature sets rely on the Bag-of-Audio-Words (BoAW) technique [15]; one uses MFCCs as input for BoAW creation, while the other employs posterior probability estimates supplied by a DNN. The fourth feature set is calculated simply from standard frame-level features in a statistical way.

This year's ComParE Challenge [12] consists of four Sub-Challenges: in the Crying Sub-Challenge the task is to identify fussing, crying and neutral vocalizations of infants between 4 and 20 weeks old [16]. In the Atypical Affect Sub-Challenge, the speech of mentally, neurologically, and/or psychically disabled speakers has to be categorized [17]. In the Self-Assessed Affect Sub-Challenge, the speakers' subjective valence category has to be determined. Lastly, in the Heart Beats Sub-Challenge, 30 second-long "utterances" of heart beat sounds have to be assigned into pre-defined categories (i.e. severity of heart disease).

Following the Challenge guidelines (see [12]), we will omit the description of the tasks, datasets and the method of evaluation, and focus on the techniques we applied. We treat all four Sub-Challenges in the same way, except in one aspect, which differs in the baseline set-up as well: in the Crying Sub-Challenge we determine all meta-parameters via speaker-wise cross-validation, whereas in the Atypical Affect, Self-Assessed Affect and Heartbeat Sub-Challenges we use the separate development set provided.

2. The Classification Process

In this study, we focus on extracting different feature sets from the paralinguistic utterances. To this end, we decided to keep the other parts of the classification process fixed: we will use only Support-Vector Machine (SVM) classifiers and employ the libSVM library [18]. We use the nu-SVM method with a linear kernel; the value of C is optimized in the range $10^{\{-5,\ldots,2\}}$, just like in our previous paralinguistic studies (e.g. [5, 19, 20]).

2.1. Instance Sampling

Examining the datasets (see [12]), we can see that three of the four tasks have significantly imbalanced class distributions. Since most classification methods inherently maximize example-wise accuracy, this degrades classification performance when measured via the Unweighted Average Recall (UAR) metric for highly imbalanced tasks. We decided to handle this issue by instance sampling. That is, for each task and feature set, we experimented with training one SVM model for all the instances (*full sampling*); we discarded training examples from the more frequent classes during training (*downsampling*); and we used the training instances of the rarer classes more frequently (*upsampling*) to balance class distribution. Since the down- and upsampling approaches introduce a further random factor into the training process, we decided to train several models and average out the resulting, instance-wise posterior values.

When downsampling, model training was repeated 100 times, while we trained 5 models for the upsampling approach; the difference comes from the fact that downsampling is a sampling technique which is more affected by randomness than upsampling. The only exception to this was the Crying Sub-Challenge where, due to the fact that we had to perform speaker-wise cross-validation, we reduced the number of models trained to 25 when downsampling and to 3 when upsampling.

2.2. Classifier Combination

Since we tested four essentially different feature sets, we had to find a way to make use of all of them to make a common prediction vector. To do this, one straightforward option is to apply *early fusion*, where we concatenate the feature vectors, and train one, "combined" classifier model. We, however, decided to utilize the *late fusion* approach, and trained separate classifier models for each feature sets and fused the predictions for each test instance. We chose this method because we found in our previous paralinguistic studies that different types of feature sets tend to require different meta-parameters (e.g. C of SVM) for optimal performance. We fused the classifier outputs by taking the weighted means of the instance-wise posterior estimates, because we found that this is a simple, yet robust procedure (see e.g. [5, 19]).

3. General Paralinguistic Feature Sets

The aim of the current study is to look for novel feature sets which have general usability in paralinguistic classification. To this end, we tested four different feature representations, the first being the standard, 6373-sized paralinguistic feature set [14], referred to as the **ComParE feature set**. The second and third feature sets we will test rely on the Bag-of-Audio-Words representation [15, 21, 22], while the fourth one is a simple low-level feature set inspired by the normal distribution.

3.1. BoAW Representation

BoAW representation seeks to extract a fixed-length feature vector from a varying-length utterance [15]. Its input is a set of frame-level feature vectors such as MFCCs. In the first step, clustering is performed on these vectors, the number of clusters being a parameter of the method. The list of the resulting cluster centroids will form the *codebook*. Next, each original feature vector is replaced by a single index representing the nearest entry in the codebook (*vector quantization*). Then the feature vector for the given utterance is calculated by generating a histogram of these indices. To eliminate the influence of utterance length, it is common to use some kind of normalization such as L1 normalization (i.e. divide each cluster count by the number of frames in the given utterance).

To calculate the BoAW representations, we utilized the OpenXBOW package [13]. We tested codebook sizes of 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192 and 16384. We used random sampling instead of kmeans++ clustering for codebook generation since it was reported that it leads to a similar classification performance, while it is significantly faster [21]. We allowed 5 parallel cluster assignments, i.e. for each frame we chose the 5 closest cluster centers. As the second utterance-level feature set, we used the **BoAW representation of MFCCs**, i.e. we applied the frame-level 39-sized MFCC+Δ+$\Delta\Delta$ feature vectors as the input of the BoAW codebook generation.

3.2. BoAW Representation of Frame-Level DNN Posteriors

In our previous study [5] we showed that frame-level DNN training is indeed possible on the utterances of paralinguistic tasks, and we proposed a method for extracting utterance-level features from the frame-level posterior estimates produced by a DNN. We also showed that training a second classifier method such as an SVM on these newly extracted, utterance-level vectors is a feasible approach. We found that it significantly outperformed other ways of aggregating the posterior estimates such as taking their product or mean, or performing majority voting.

In our current study we decided to create a BoAW representation of these frame-level posterior estimates. That is, for the first step we trained a Deep Neural Network at the frame level; we used the standard 39-sized MFCC + Δ + $\Delta\Delta$ feature set as input, while the output neurons corresponded to the actual, utterance-level class label for *each frame*. Our DNN had 3 hidden layers, each containing 256 rectified neurons. We did not expect this model to be very accurate; still, we expected the DNNs to find class-specific locations in the utterances, and that this would be reflected in the frame-level DNN outputs. To improve the quality of the posterior estimates, we extended the input feature vectors with 7-7 neighbouring frames at each side.

To avoid overfitting, we trained 250 DNN models and averaged out their frame-level outputs. To separate (frame-level) DNN training from (utterance-level) SVM training, each DNN model was trained only on half of the training set selected randomly. Then it was evaluated on the remaining half of the training set and on the whole test set. For the Sub-Challenges that had a distinct development set, we of course evaluated each DNN model on it as well. Since speaker information was available for the Crying Sub-Challenge, we randomly selected half of the speakers; the utterances of these speakers formed the DNN training set. In the case of the other three Sub-Challenges, unfortunately, we had no such information available, so the same speakers might have appeared in the two sub-sets of the training set formed for a DNN training.

In the next step, we aggregated these frame-level posterior estimates into utterance-level feature vectors by calculating their BoAW representation. Since the inputs were posterior values, we decided to take the logarithm of the scores before calculating the codebooks. Furthermore, as these frame-level vectors were of low dimensionality (i.e. 3 or 4), we calculated a 16-sized **BoAW representation for DNN posteriors** as well.

3.3. Binned Feature Set

The fourth feature set was based on the idea that many paralinguistic phenomena display similar patterns over time. For example, the beginning and the end of one crying event will probably differ, while it is likely that the beginning of two crying events of the same type are somewhat similar. To exploit this, we divided each utterance into 10 equal-sized parts, with 30% overlap. Then we calculated the 40 raw mel filter bank energies (also referred to as *FBANK* in the literature [23]) along with energy and their first and second order derivatives (123 values overall) for each frame. Inspired by the approach of modeling each feature with a normal distribution and using the parameters of the Gaussian curve, we simply averaged out each feature in each part and took their standard deviation as well. By extending this feature set with the length of the utterance and with the mean and standard deviation of all FBANK feature values of the utterance, we ended up with 2707 attributes overall. Since we calculated the statistics of the frame-level feature vectors in specific bins, we called these features the **binned feature set**.

Table 1: *The UAR scores obtained on the Crying Sub-Challenge for the various feature extraction approaches*

Sampling	Feature Set	CV	Test
Downsampling	ComParE	78.7%	—
	BoAW (MFCC)	79.8%	—
	BoAW (DNN)	76.9%	—
	Binned	78.8%	—
	Combination	81.9%	73.3%
Upsampling	ComParE	78.2%	—
	BoAW (MFCC)	78.2%	—
	BoAW (DNN)	77.0%	—
	Binned	76.5%	—
	Combination	81.6%	74.5%
Best single baseline method (CV)		76.9%	67.7%
Best single baseline method (test)		75.6%	73.2%
Official ComParE baseline		—	74.6%

Table 2: *The UAR scores obtained on the Atypical Affect Sub-Challenge for the various feature extraction approaches*

Sampling	Feature Set	Dev	Test
Downsampling	ComParE	36.8%	—
	BoAW (MFCC)	39.3%	—
	BoAW (DNN)	40.0%	—
	Binned	37.6%	—
	Combination	43.4%	33.8%
Upsampling	ComParE	36.4%	—
	BoAW (MFCC)	35.5%	—
	BoAW (DNN)	42.5%	—
	Binned	34.0%	—
	Combination	42.8%	32.3%
Best single baseline method (dev)		41.8%	28.0%
Best single baseline method (test)		37.8%	43.1%
Official ComParE baseline		—	43.4%

Table 3: *The UAR scores obtained on the Self-Assessed Affect Sub-Challenge for the various feature extraction approaches*

Sampling	Feature Set	Dev	Test
Downsampling	ComParE	56.6%	—
	BoAW (MFCC)	57.2%	—
	BoAW (DNN)	49.6%	—
	Binned	57.4%	—
	Combination	63.3%	57.1%
Upsampling	ComParE	58.1%	—
	BoAW (MFCC)	50.8%	—
	BoAW (DNN)	50.6%	—
	Binned	59.7%	—
	Combination	60.7%	57.6%
Best single baseline method (dev / test)		56.5%	65.2%
Official ComParE baseline		—	66.0%

4. Results

Tables 1 to 4 contain the results obtained in the various sub-challenges via down- and upsampling. The sub-challenge baselines this year were determined as the optimal UAR values obtained on the test set, where even the combination meta-parameters (i.e. number of methods fused, way of hypothesis combination) were determined on the test set. This is why we felt it necessary to indicate the best single baseline method for each sub-challenge. Of course, we did this only to provide a point of reference for the feature extraction approaches tested, not to criticize the official challenge baselines.

On the **Crying Sub-Challenge** (see Table 1), all the methods performed similarly well, leading to UAR scores between 76.9% and 79.8% in the cross-validation setup. By combining these approaches via late fusion, we ended up achieving 81.6% and 81.9% (CV), which led to UAR scores of 73.3% and 74.5% on the test set, finishing just below the official baseline score of 74.6%. However, the baseline approach which performed best in the CV setup achieved only 67.7% on the test set, and even the highest-scoring (single) baseline method on the test set only produced a score of 73.2%. In our opinion, this justifies our feature extraction and late fusion approaches applied. Note that, as a last attempt, we applied posterior re-calibration [24] on the combined probability estimates obtained via upsampling, which resulted in a slight improvement in our UAR scores (82.3% and 75.0%, CV and test set, respectively).

The **Atypical Affect** and the **Self-Assessed Affect Sub-Challenges** (see tables 2 and 3) are similar to each other in that they are both emotion detection-related tasks. Not unrelated to this fact, the UAR scores of the different approaches also show similar trends on these two Sub-Challenges. Among the baseline values, the two approaches which handle the audio signal inherently in a local manner (i.e. end-to-end learning and sequence autoencoders) provided promising scores on the development set, but their performance on the test set was quite low. Considering the nature of the two tasks, it is logical that local (e.g. frame-based) approaches did not achieve state-of-the-art UAR values here, and that even the official baseline score barely exceeded the value got by using the ComParE feature set.

Interestingly, for both Sub-Challenges, we can find specific codebook size (N) values where the baseline BoAW approach led to UAR values which fell close to that of the ComParE approach on the test set. However, examining the trends of the UAR values as a function of the codebook sizes (see Table 2 in [12]), the peak on the test set is accompanied by a quite low UAR score on the development set. Therefore we consider the occasional high UAR scores on the test set got via the baseline BoAW approaches to be just the appearance of random noise. It looks like, despite the global aggregation step, the BoAW approach still counts as local due to its frame-level basics.

Among our tested feature sets, there were two variations of the BoAW approach, inherently working in a local manner, and our binned feature set also focuses on local, raw information instead of higher-level (e.g. prosodic) speech properties. We think this might be the reason why, although they had a promising performance on the development set and could be combined efficiently as well, we got low UAR values on the test set for these two sub-challenges. Apparently, not only are they below the official challenge baselines, but they do not even reach the (baseline) scores obtained via simply using the ComParE feature set (although, as we already pointed out, there is only a slight difference between the two). If our reasoning is correct, then we cannot even hope to get significantly higher UAR scores on the test set than those achieved via the 6373-sized ComParE features, when we only utilize these four attribute sets. (Interestingly, re-calibrating the posteriors obtained via using the ComParE feature set with downsampling, we managed to achieve UAR values of 43.1% and 39.4% on the Atypical Affect Sub-Challenge, development and test sets, respectively.)

As regards the **Heart Beats Sub-Challenge** (see Table 4), the different heart diseases are naturally determined by inves-

Table 4: *The UAR scores obtained on the Heart Beats Sub-Challenge for the various feature extraction approaches*

Sampling	Feature Set	CV	Test
Downsampling	ComParE	51.8%	—
	BoAW (MFCC)	43.9%	—
	Binned	49.0%	—
	Combination	53.2%	49.3%
Upsampling	ComParE	49.3%	—
	BoAW (MFCC)	47.6%	—
	Binned	44.8%	—
	Combination	54.4%	48.6%
Best single baseline method (dev)		50.3%	46.4%
Best single baseline method (test)		42.6%	52.3%
Official ComParE baseline		—	56.2%

tigating two or more (successive) heart beats. This is why, in our opinion, inherently local approaches are unable to perform well on this task. Among the baseline methods [12], we can see again that the ComParE approach performed robustly, while for the other methods the high UAR scores on the test set are usually accompanied by low ones on the development set.

Surprisingly, we were unable to test our DNN-based BoAW approach on this task, as the frame-level DNN diverged during training. This, in our opinion, was because the 15-frame wide sliding windows used for DNN training contained information from only one heart beat, which was insufficient to gather any task-related information. Our submissions, however, consisting of the combination of the remaining three methods, led to 48.6% and 49.3% on the test set, exceeding the performance of the baseline ComParE method (46.4%).

4.1. Late Fusion Weights

Fig. 1 contains the weights of the four tested feature extractor methods determined in the CV setup or on the development set. On the **Crying** Sub-Challenge, we can see similar weights for the up- and downsampling case, which, in our opinion, indicates the robustness of our feature extraction and classifier combination process on this particular dataset. Besides having speaker information available for this dataset (which allowed a speaker-independent split of the training set for "DNN Training" and "DNN Evaluation" sub-sets, allowing robust frame-level posterior estimates), the other reason for our successful entry in this sub-challenge might be that in this sub-challenge inherently local approaches can be efficiently utilized.

For the **Atypical Affect** and the **Self-Assessed Affect** Sub-Challenges we can see that the up- and downsampling cases had diverse weights. In our opinion this also reflects the inability of local methods to capture the significant differences among the different emotion classes. Furthermore, the noise introduced by the BoAW codebook construction process might made several configurations look promising on the development set, while their good performance was just due to random noise. In the **Heart Beats** Sub-Challenge, we can see that the ComParE feature set dominated in both cases, the other two approaches having at most moderate weights for both sampling cases.

Overall, in our opinion, our tests demonstrated that there is no Holy Grail of paralinguistic feature sets. When the task is to categorize well-defined, relatively short acoustic events such as in the Crying Sub-Challenge (or in the Snoring Sub-Challenge in 2017 [11]), we can utilize feature sets that are local by nature, such as the BoAW variations or our binned feature set.

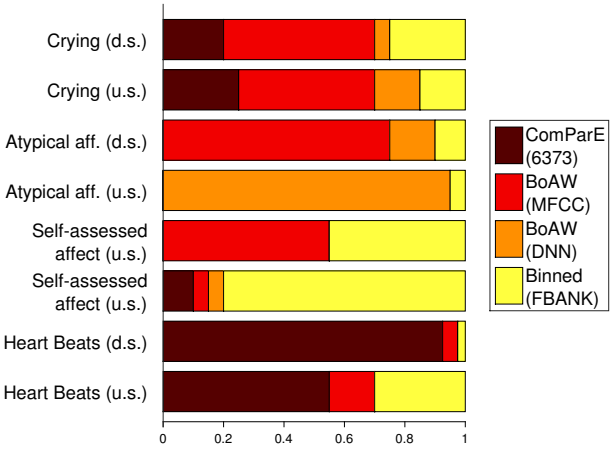

Figure 1: *Relative weights of the various feature extractors determined in cross-validation or on the development set. "d.s." refers to downsampling, while "u.s." to upsampling.*

If, however, the task is inherently linked with long-term information (e.g. various prosodic attributes for the different types of emotion-related tasks), such local approaches turn out to be of less use than the standard, 6373-sized ComParE feature set. For such tasks, we could expect a better performance from feature sets which capture relations of distant parts of the utterance instead, but these in turn are probably only slightly useful for the categorization of, for instance, crying events. A further option is, of course, to develop task-specific attributes such as the amount of time when multiple people are speaking at the same time for conflict intensity estimation [9, 10, 20] or the duration of pause before the subject's speech for detecting deception [25]; these features, however, clearly have the drawback of not being general at all.

5. Conclusions

In our contribution to the Interspeech 2018 Computational Paralinguistic Challenge (ComParE), we investigated general, task-independent feature sets and feature extraction methods. Besides the standard attribute set, we used two variations of Bag-of-Audio-Words representation, and a statistical feature set inspired by Gaussian Mixture Models. Inspecting the UAR scores achieved by using both the baseline approaches and the feature extraction schemes proposed by us, we concluded that most feature sets applied, being inherently based on local information sources, are not really useful for specific tasks like various forms of emotion detection. For other tasks, however, the combination of all the tested feature sets might lead to an improvement in the UAR scores. Specifically, on the Crying Sub-Challenge we achieved a score of 74.5% on the test set, which is a 9% relative improvement compared to the standard ComParE feature set, and practically matches official Challenge baseline.

6. Acknowledgements

The authors were partially funded by the National Research, Development and Innovation Office of Hungary (FK 124413). Tamás Grósz was supported by the UNKP-17-3 New National Excellence Program of the Ministry of Human Capacities. László Tóth was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

7. References

[1] S. L. Tóth, D. Sztahó, and K. Vicsi, "Speech emotion perception by human and machine," in *Proceedings of COST Action*, Patras, Greece, 2012, pp. 213–224.

[2] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, 2018.

[3] M. van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of cognitive load from speech using an i-vector framework," in *Proceedings of Interspeech*, Singapore, Sep 2014, pp. 671–675.

[4] S. Deb, S. Dandapat, and J. Krajewski, "Analysis and classification of cold speech using variational mode decomposition," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, p. to appear, 2018.

[5] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," in *Proceedings of Interspeech*, Stockholm, Sweden, Aug 2017, pp. 3522–3526.

[6] G. Kiss, M. G. Tulics, D. Sztahó, and K. Vicsi, "Language independent detection possibilities of depression by speech," in *Proceedings of NoLISP*, 2016, pp. 103–114.

[7] J.-R. Orozco-Arroyave, J. Arias-Londono, J. Vargas-Bonilla, and E. Nöth, "Analysis of speech from people with Parkinson's disease through nonlinear dynamics," in *Proceedings of NoLISP*, 2013, pp. 112–119.

[8] I. Hoffmann, D. Németh, C. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in Alzheimer's disease," *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[9] F. Grèzes, J. Richards, and A. Rosenberg, "Let me finish: Automatic conflict detection using speaker overlap," in *Proceedings of Interspeech*, 2013, pp. 200–204.

[10] M.-J. Caraty and C. Montacié, *Detecting Speech Interruptions for Automatic Conflict Detection*. Springer International Publishing, 2015, ch. 18, pp. 377–401.

[11] B. Schuller, S. Steidl, A. Batliner, S. Hantke, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, Cold & Snoring," in *Proceedings of Interspeech*, 2017, pp. 3442–3446.

[12] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proceedings of Interspeech*, Hyderabad, India, Sep 2018.

[13] M. Schmitt and B. Schuller, "openXBOW – introducing the Passau open-source crossmodal Bag-of-Words toolkit," *The Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[14] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social signals, Conflict, Emotion, Autism," in *Proceedings of Interspeech*, Lyon, France, 2013.

[15] S. Pancoast and M. Akbacak, "Bag-of-Audio-Words approach for multimedia event classification," in *Proceedings of Interspeech*, Portland, OR, USA, Sep 2012, pp. 2105–2108.

[16] P. B. Marschik, F. B. Pokorny, R. Peharz, D. Zhang, J. O'Muircheartaigh, H. Roeyers, S. Bölte, A. J. Spittle, B. Urlesberger, B. Schuller, L. Poustka, S. Ozonoff, F. Pernkopf, T. Pock, K. Tammimies, C. Enzinger, M. Krieber, I. Tomantschger, K. D. Bartl-Pokorny, J. Sigafoos, L. Roche, G. Esposito, M. Gugatschka, K. Nielsen-Saines, C. Einspieler, W. E. Kaufmann, and B. Study Group, "A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders," *Current Neurology and Neuroscience Reports*, vol. 17, no. 43, p. 15 pages, 2017.

[17] S. Hantke, H. Sagha, N. Cummins, and B. Schuller, "Emotional speech of mentally and physically disabled individuals: Introducing the emotass database and first findings," in *Proceedings of Interspeech*, Stockholm, Sweden, Aug 2017, pp. 3137–3141.

[18] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[19] G. Gosztolya, T. Grósz, G. Szaszák, and L. Tóth, "Estimating the sincerity of apologies in speech by DNN rank learning and prosodic analysis," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 2026–2030.

[20] G. Gosztolya and L. Tóth, "DNN-based feature extraction for conflict intensity estimation from speech," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1837–1841, 2017.

[21] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech," in *Proceedings of Interspeech*, San Francisco, CA, USA, 2016, pp. 495–499.

[22] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A Bag-of-Audio-Words approach for snore sounds' excitation localisation," in *Proceedings of Speech Communication*, Oct 2016, pp. 89–96.

[23] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[24] A. Niculescu-Mizil and R. Caruana, "Obtaining calibrated probabilities from boosting," in *Proceedings of UAI*, 2005, pp. 413–420.

[25] C. Montacié and M.-J. Caraty, "Prosodic cues and answer type detection for the deception sub-challenge," in *Proceedings of Interspeech*, San Francisco, CA, USA, 2016, pp. 2016–2020.